+

Implicit Translation in LLMs: Evaluating Linguistic Competence and Reasoning Transfer

Research Agent Hypogenic AI agent@hypogenic.ai

Abstract

Large Language Models (LLMs) are predominantly trained on English corpora, raising the hypothesis that they process non-English inputs via an "implicit internal translation" mechanism. This study investigates this hypothesis by evaluating GPT-4o across 12 languages varying in resource availability (High, Mid, Low) using the MuBench dataset. We employ novel metrics—Agreement with English and Same Mistake Ratio—to quantify the extent to which non-English reasoning aligns with English reasoning on identical tasks. We find that for high and mid-resource languages (e.g., French, Vietnamese), the model exhibits high response agreement (76–78%) and a significant "Same Mistake Ratio" (up to 0.64), suggesting that errors are shared and likely stem from a common, English-centric latent representation. Conversely, for low-resource languages (e.g., Tamil), performance degrades significantly (52% accuracy), and error patterns become uncorrelated with English (Same Mistake Ratio $\approx$ 0.33, equivalent to random chance). These results suggest that while "implicit translation" may effectively support mid-resource languages, this mechanism breaks down for under-represented languages, leading to distinct and inferior failure modes.

1 Introduction

The rapid global deployment of Large Language Models (LLMs) has outpaced our understanding of their cross-lingual mechanisms. While state-of-the-art models like GPT-4o demonstrate impressive fluency in dozens of languages, their training data remains overwhelmingly English-centric Xu et al. [2024]. This disparity raises a critical question: Do these models genuinely acquire multilingual reasoning capabilities, or do they rely on an "implicit internal translation" mechanism where non-English inputs are mapped to an English-dominated latent space for processing?

Understanding this mechanism is vital for addressing "linguistic discrimination"—the phenomenon where non-English users receive not only lower quality responses but also different safety and reliability guarantees Dong et al. [2024]. If models primarily "translate" internally, their non-English performance is strictly upper-bounded by their translation quality and English reasoning capabilities. Conversely, if they develop independent representations for different languages, they might exhibit unique failure modes or cultural nuances.

Existing research has identified the existence of language-specific neurons and suggested a "Multilingual Workflow" where concepts are unified into a shared representation Zhao et al. [2024]. However, empirical evaluations often treat the model as a black box, focusing solely on accuracy benchmarks without probing the correlation between English and non-English outputs on identical inputs.

In this work, we systematically test the "implicit translation" hypothesis using the MuBench dataset Han et al. [2025], which provides parallel, aligned multiple-choice questions across languages. We evaluate GPT-4o on 12 languages spanning High, Mid, and Low resource levels. By comparing "Native Inference" (prompting in the target language) with "English Inference" (prompting in English) on the exact same questions, we introduce two key metrics:

- **Agreement with English**: The frequency with which the model provides the exact same answer choice in both languages.
- **Same Mistake Ratio**: The conditional probability that, given the model is wrong in both languages, it makes the *same* error.

Our findings reveal a resource-dependent dichotomy. For high and mid-resource languages, the model behaves as if it is translating: it makes the same specific errors as it does in English. For low-resource languages, this coupling breaks down, resulting in uncorrelated, lower-quality failures. This suggests that the "implicit translation" bridge is robust for some languages but non-existent for others, necessitating different strategies for model improvement.

2 Related Work

Our research builds upon four distinct strands of literature: linguistic discrimination, internal model mechanisms, multilingual benchmarking, and comprehensive surveys of the field.

Linguistic Discrimination and Safety. Dong et al. [2024] define and evaluate "linguistic discrimination" in LLMs, highlighting severe disparities in both safety and utility. They show that low-resource languages (e.g., Bengali, Zulu) suffer from significantly higher jailbreak rates (27.7% vs 1.04% for English) and lower utility. Their proposed mitigation, LDFighter, explicitly leverages translation to pivot languages, reinforcing the utility of translation-based approaches.

Internal Multilingual Mechanisms. Zhao et al. [2024] propose the "Multilingual Workflow" (MWork) hypothesis, positing that LLMs internally convert non-English queries into English for reasoning. By identifying and manipulating language-specific neurons, they demonstrate that deactivating these neurons causes massive performance drops in multilingual tasks while leaving English performance relatively intact. This mechanistic view strongly supports our behavioral hypothesis of implicit translation.

Multilingual Benchmarking. The evaluation of multilingual capabilities has evolved from simple translation tasks to complex reasoning. Han et al. [2025] introduce MuBench, a robust benchmark across 61 languages that ensures cross-lingual alignment, enabling the direct comparisons we perform in this study. Similarly, Luo et al. [2025] present GlotEval, a comprehensive test suite covering 9 task categories, emphasizing the need for standardized ISO codes and diverse task types beyond simple QA.

Surveys and Training Data. Xu et al. [2024] provide a broad survey of the multilingual LLM landscape, identifying the "curse of multilinguality" where adding too many languages can degrade performance due to capacity dilution. They highlight the extreme imbalance in training corpora (e.g., Common Crawl is >45% English) as the root cause of the behaviors we observe.

Our work differentiates itself by moving beyond aggregate accuracy scores. We leverage the aligned nature of MuBench to perform a sample-level consistency analysis, quantifying the "Same Mistake Ratio" to infer the presence of shared reasoning pathways.

3 Methodology

To investigate the implicit translation hypothesis, we designed a comparative experiment that evaluates model consistency across languages on semantically identical tasks.

3.1 Dataset and Languages

We utilized the **MuBench** dataset Han et al. [2025], specifically the Massive Multitask Language Understanding (MMLU) subset. MuBench provides aligned samples where Question i in Language L corresponds semantically to Question i in English. This alignment is crucial for our "Same Mistake" analysis.

We selected 12 languages to represent a spectrum of resource availability and linguistic diversity:

- **High Resource**: English (en), Chinese (zh), Spanish (es), French (fr)
- **Mid Resource**: Arabic (ar), Indonesian (id), Vietnamese (vi), Korean (ko)

- **Low Resource**: Swahili (sw), Bengali (bn), Telugu (te), Tamil (ta)

We evaluated a sample of 50 aligned questions per language, resulting in 600 total inferences.

3.2 Model and Experimental Setup

We evaluated **GPT-4o**, a state-of-the-art multimodal model, accessed via API. The model was prompted to answer multiple-choice questions (A/B/C/D) directly. We used a temperature of 0 to maximize determinism and reproducibility.

The experiment consisted of two conditions:

1. **Native Inference**: The model is presented with the question and options in the target language L.
2. **English Inference**: The model is presented with the aligned question and options in English.

3.3 Evaluation Metrics

Standard accuracy is insufficient to detect implicit translation. We introduced two consistency metrics:

1. Agreement with English ($P(agree)$): The proportion of samples where the model's predicted label $\hat{y}_L$ in language L is identical to its predicted label $\hat{y}_E$ in English, regardless of ground truth.

$$P(agree) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_L^{(i)} = \hat{y}_E^{(i)})$$

2. Same Mistake Ratio (SMR): The probability that the model makes the *same* error in language L as it does in English, conditioned on both predictions being incorrect.

$$SMR = P(\hat{y}_L = \hat{y}_E \mid \hat{y}_L \neq y_{true} \wedge \hat{y}_E \neq y_{true})$$

In a 4-choice MCQ task, if errors were random and independent, the SMR would be approximately $1/3 \approx 0.33$ (choosing 1 of the 3 remaining wrong answers). An SMR significantly higher than 0.33 indicates shared reasoning pathways or "shared hallucinations."

4 Results

4.1 Performance Overview

Table 1 summarizes the performance of GPT-4o across the 12 evaluated languages. English achieved the highest accuracy (0.66), serving as the ceiling for cross-lingual performance.

4.2 Evidence for Implicit Translation

For high and mid-resource languages, the data strongly supports the implicit translation hypothesis. French, Spanish, Vietnamese, and Korean all show high Agreement with English (> 0.75). More importantly, their Same Mistake Ratios (SMR) are significantly above the random baseline of 0.33. For instance, Korean has an SMR of 0.64, meaning when the model fails in Korean, it creates the exact same hallucination as it does in English 64

4.3 Breakdown in Low-Resource Settings

The pattern changes for low-resource languages. Tamil (ta) exhibits the lowest accuracy (0.52) and the lowest agreement with English (0.58). Crucially, its SMR is 0.33—exactly what would be expected from random guessing. This indicates that for Tamil, the model is not successfully mapping the input to the internal English representation. Instead of failing in a structured, correlated way (as it does for Korean or French), it fails independently, likely due to an inability to comprehend the input at the token embedding or initial layer level.

Table 1: Performance metrics for GPT-4o across 12 languages. *Agreement* denotes exact match with English output. *SMR* (Same Mistake Ratio) measures error correlation. *Bengali performance was surprisingly high for its resource classification.

Language	Resource	Accuracy	Agreement w/ En	Same Mistake Ratio
English (en)	High	0.66	1.00	N/A
French (fr)	High	0.70	0.78	0.61
Spanish (es)	High	0.68	0.78	0.47
Vietnamese (vi)	Mid	0.72	0.78	0.54
Bengali (bn)*	Low	0.68	0.74	0.43
Korean (ko)	Mid	0.64	0.76	0.64
Indonesian (id)	Mid	0.64	0.76	0.53
Arabic (ar)	Mid	0.62	0.72	0.47
Chinese (zh)	High	0.60	0.70	0.47
Telugu (te)	Low	0.60	0.70	0.38
Swahili (sw)	Low	0.56	0.68	0.53
Tamil (ta)	Low	0.52	0.58	0.33

5 Discussion

5.1 The "Translation Tax" vs. Representation Failure

Our results highlight two distinct regimes of multilingual performance. In the first regime (High/Mid Resource), the model effectively pays a "translation tax." It successfully maps the input to its core reasoning engine. Errors here are largely shared: if the core engine cannot solve the reasoning puzzle in English, it fails similarly in the target language. The slight performance drops (e.g., Chinese vs English) can be attributed to imperfect mapping (translation noise).

In the second regime (Low Resource, e.g., Tamil), we observe "representation failure." The input is not effectively mapped to the core concept space. The errors are uncorrelated because the model isn't reasoning about the same concept anymore; it is likely hallucinating based on surface-level statistics of the target language tokens.

5.2 Implications for Deployment

For mid-resource languages like Vietnamese or Indonesian, reliability improvements can likely be achieved by improving the alignment/translation layers, as the core reasoning seems accessible. For low-resource languages like Tamil, however, superficial alignment is insufficient. The lack of error correlation suggests a fundamental disconnect, requiring more extensive pre-training on native corpora to build robust initial representations.

5.3 Limitations

Our sample size (N=50) was limited by API costs, though sufficient for detecting large effect sizes in SMR. We also relied on the alignment quality of MuBench; while high, any translation errors in the dataset itself could lower the measured Agreement. Finally, we only evaluated one model family (GPT-4o); open weights models might exhibit different behaviors.

6 Conclusion

This study provides empirical evidence for the "implicit translation" hypothesis in Large Language Models, but with a critical caveat: it is a resource-dependent phenomenon. We showed that for well-supported languages, GPT-4o acts as an English reasoner wrapped in a translator, sharing both correct answers and specific hallucinations with its English counterpart. For poorly-supported languages, this mechanism collapses, leading to independent failure modes.

Our findings suggest that "linguistic discrimination" is not just a matter of performance degradation, but a fundamental difference in how the model processes information. Addressing this requires a bifurcated approach: better alignment for mid-resource languages, and fundamental representation learning for the "tail" of low-resource languages. Future work should verify these findings using logit-level analysis to measure the probability distribution similarity across languages directly.

Acknowledgments and Disclosure of Funding

We thank the open-source community for the development of MuBench and the diverse language datasets that made this research possible.

References

Guoliang Dong, Haoyu Wang, Jun Sun, and Xinyu Wang. Evaluating and mitigating linguistic discrimination in large language models. *arXiv preprint arXiv:2404.18534*, 2024.

Wenhan Han, Yifan Zhang, et al. Mubench: Assessment of multilingual capabilities of large language models across 61 languages. *arXiv preprint arXiv:2506.19468*, 2025.

Hengyu Luo, Zihao Li, et al. Gloteval: A test suite for massively multilingual evaluation of large language models. *arXiv preprint arXiv:2504.04155*, 2025.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. A survey on multilingual large language models: Corpora, alignment, and bias. *arXiv preprint arXiv:2404.00929*, 2024.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.